TD4

Issues de l'enquête Formation des couples menée par l'institut national des études démographiques dans les années 1980, les données analysées croisent le groupe social des hommes avec le lieu où ils ont rencontré leur conjointe. À l'aide des résultats ci-dessous, répondez aux questions suivantes :

1. Décrivez la base de données étudiée.
2. Réalisez l'analyse en justifiant l'usage de la méthode d'analyse géométrique des données à utiliser ici.
3. Combien de valeurs propres faut-il retenir ? Justifez.
4. Interprétez et commentez les différents axes.

Préparation de l'analyse

D'abord, comme d'habitude, il faut indiquer le bon répertoire de travail où se trouve la base données que nous utilisons.

```
# Répertoire de travail
setwd("E:/Enseignements/Cours UVSQ/Année 2023-2024/L3 Sociologie Analyse des données/TD/TD
```

Dans ce TD, nous utilisons les packages listés ci-dessous.

Le package readr permet d'ouvrir la base de données.

Le package tibble permet de faire certains recodages de la base de données facilement.

FactoMineR est le package permettant d'utiliser les fonctions d'analyse géométrique des données (ACP, AFC…). Le package explor permet de lire facilement les résultats d'une analyse. Le package factoextra contient plusieurs fonctions utiles pour analyser les résultats d'une analyse.

```r
# Appel des packages
library(readr)
library(tibble)
library(FactoMineR)
library(explor)
library(factoextra)
```

```
Le chargement a nécessité le package : ggplot2
```

```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

Utilisons la fonction read_csv2 pour charger la base de données qui s'appelle "TD4.csv". R nous dit à cette occasion qu'il y a 16 lignes et 15 colonnes dans la base, qu'une variable (PCS) est de format "chr", donc catégorielle, les autres sont de format "dbl" donc numériques.

La ligne de code suivante utilise la fonction column_to_rownames du package tibble pour indiquer que la variable (colonne) PCS doit être considérée comme le libellé des lignes de la base.

```r
setwd("E:/Enseignements/Cours UVSQ/Année 2023-2024/L3 Sociologie Analyse des données/TD/TD
# Lecture du fichier : on place la première colonne comme nom de lignes
TD4 <- read_csv2("TD4.csv")
```

```
i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
Rows: 16 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ";"
chr  (1): PCS
dbl (14): Etudes, Vacances, Fete entre amis, Association, Travail, Particuli...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#Positionner la première colonne (nom des PCS) comme libellé des lignes (plutôt que comme
TD4<-column_to_rownames(TD4, var = "PCS")
```

On peut avoir une d'à quoi ressemble plus précisément la base avec la fonction head ou en tapant View(TD4) ce qui va ouvrir un onglet avec la base.

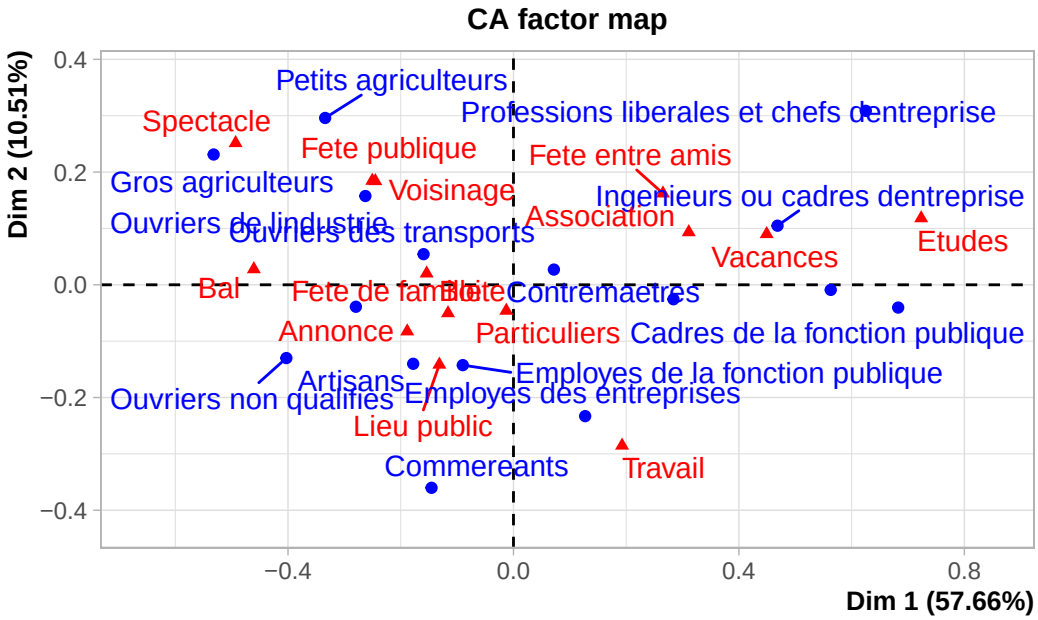
```
head(TD4)
```

Conduite de l'AFC

La base est déjà prête pour l'analyse : il s'agit bien d'un tableau de contingence qui croise en ligne le groupe social des hommes (PCS) avec en colonne le lieu de rencontres du partenaire. Dans chaque cellule se trouve les effectifs. Ces données se prêtent bien à une AFC.

On conduit une AFC grâce à la fonction CA du package FactoMineR sur la base TD4. Les résultats de l'analyse sont stockés dans l'objet res.afc. La fonction CA nous sort automatiquement des graphiques.

```
# Réalisation de l'AFC
res.afc<-CA(TD4)
```



Analyse des valeurs propres

D'abord, on va déterminer combien d'axes à retenir dans notre analyse. Ici, il y a 16 lignes et 14 colonnes dans notre tableau de contingence, donc au maximum on peut interpréter min(16,14)-1=13 axes. Bien sur, c'est beaucoup trop, alors on regarde les valeurs propres

associées à chaque axe, correspondant à l'inertie expliquée. En AFC, l'inertie du nuage des profils-lignes et des profils-colonnes correspond à l'écart à l'indépendance des données observées par rapport à la situation d'indépendance si le lieu de rencontre du partenaire ne dépendait pas du groupe social des hommes.

Notons que la somme des valeurs propres sur les 13 axes, correspondant à l'inertie totale (correspondant au ϕ^2, où χ^2/n), est égale à 0.21. Ainsi, si on réalise un test du khi-deux sur notre tableau de contingence, on voit que la valeur du χ^2 est 2017,8, si on divise cette valeur par l'effectif total du tableau de contingence on retrouve bien le ϕ^2, somme des valeurs propres de l'AFC. D'ailleurs, ce test nous montre, puisque la p-value est très proche de 0, qu'il y a une association entre le groupe social des hommes et le lieu de rencontre du partenaire (nous reverrons ce test ultérieurement si vous n'en avez pas connaissance). Autrement dit, on a bien raison de conduire une AFC : là où le test du khi-deux étudie s'il y a une association ou non entre les deux variables (en comparant les observations du tableau par rapport à une situation d'indépendance), l'AFC permet d'étudier précisément le type d'association qui existe entre les deux variables.

Un dernier point : les inerties sur chacun des axes dans une AFC sont toujours comprises entre 0 et 1 (ce n'est pas le cas en ACP puisque la valeur propre est toujours supérieure ou égale à 1). Si, dans le tableau de contingence, on avait une association quasiment exclusive entre un bloc de modalités en lignes et un bloc de modalités en colonnes (on observerait des 0 dans certaines cellules du tableau), alors la valeur propre du premier axe aurait été égale à 1. Ici, ce n'est pas le cas, la valeur propre du premier axe est de 0,12, autrement dit il n'y a pas d'association exclusive entre le groupe social et le lieu de rencontre : même s'il y a association entre les deux variables, elle n'est pas parfaite.

Regardons le critère de Kaiser sur le nombre d'axes à retenir. Ce critère nous dit qu'on retient le nombre d'axes qui a une contribution supérieure à la contribution moyenne. Ici, on voit que suivant ce critère, on peut retenir 3 axes.

```
# Affichage des valeurs propres
vp<-res.afc$eig
# On ne conserve que les valeurs propres en valeur aboslue
vp[,1]
```

```
      dim 1        dim 2        dim 3        dim 4        dim 5        dim 6
1.202539e-01 2.192937e-02 1.683896e-02 1.395141e-02 1.116155e-02 9.234338e-03
      dim 7        dim 8        dim 9       dim 10       dim 11       dim 12
6.216746e-03 3.401748e-03 2.784087e-03 2.048523e-03 5.489817e-04 1.856275e-04
     dim 13
4.852756e-07
```

```
sum(vp[,1])
```

```
[1] 0.2085557
```

```r
chisq.test(TD4)
```

```
    Pearson's Chi-squared test

data:  TD4
X-squared = 2017.8, df = 195, p-value < 2.2e-16
```

```r
# On regarde lesquelles sont supérieures à l'inertie moyenne
vp[,1]>mean(vp[,1])
```

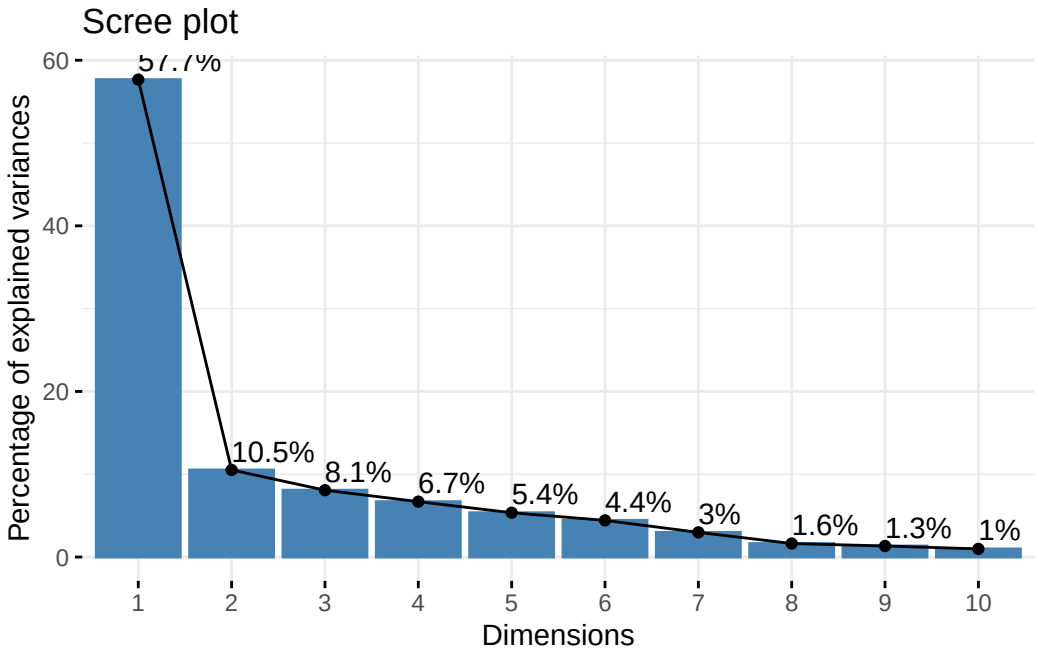
```
 dim 1  dim 2  dim 3  dim 4  dim 5  dim 6  dim 7  dim 8  dim 9 dim 10 dim 11
  TRUE   TRUE   TRUE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE  FALSE
dim 12 dim 13
 FALSE  FALSE
```

Regardons le critère du coude, où on retient le nombre d'axes avant qu'il n'y ait un éboulis trop important dans les données. Une fonction intéressante pour regarde cela est la fonction fviz_contrib du package factoextra.

L'éboulis du pourcentage de l'inertie expliquée sur chaque axe (on a pris chaque valeur propre divisée par l'inertie totale) nous montre que le premier axe permet d'interpréter 58% de l'inertie de l'AFC, soit plus de la moitié. Les deuxième et troisième axes ne contribue à expliquer que 11% et 8% de l'inertie totale.

Donc, si on peut chercher à interpréter les trois axes, il faut bien se dire que c'est surtout le premier axe qui montre les associations et les oppositions les plus structurantes dans le tableau de contingence. Les deuxième et troisième axes sont secondaires dans l'analyse.

```r
##Visualiser l'inertie du nuage projeté sur chaque axe
fviz_eig(res.afc,addlabels=T)
```



Interprétation des axes

Contribution des profils-lignes et des profils-colonnes sur chacun des axes

On peut jeter un oeil sur les représentations géométriques mais il est aussi sage de regarder quels sont les profils-lignes et les profils-colonnes qui ont les plus fortes contributions sur chacun des axes. Ces profils qui ont une contribution supérieure à la contribution moyenne seront interprétés en priorité. En effet, les modalités des lignes et des colonnes ayant influencé le plus à la construction des axes sont celles dont les contributions sont les plus élevées. On pourra se contenter d'interpréter les résultats des modalités pour lesquelles les contributions sont supérieures à la contribution moyenne (cas où chaque profil-ligne ou profil-colonne contribuerait de manière égale à la structuration de l'axe).

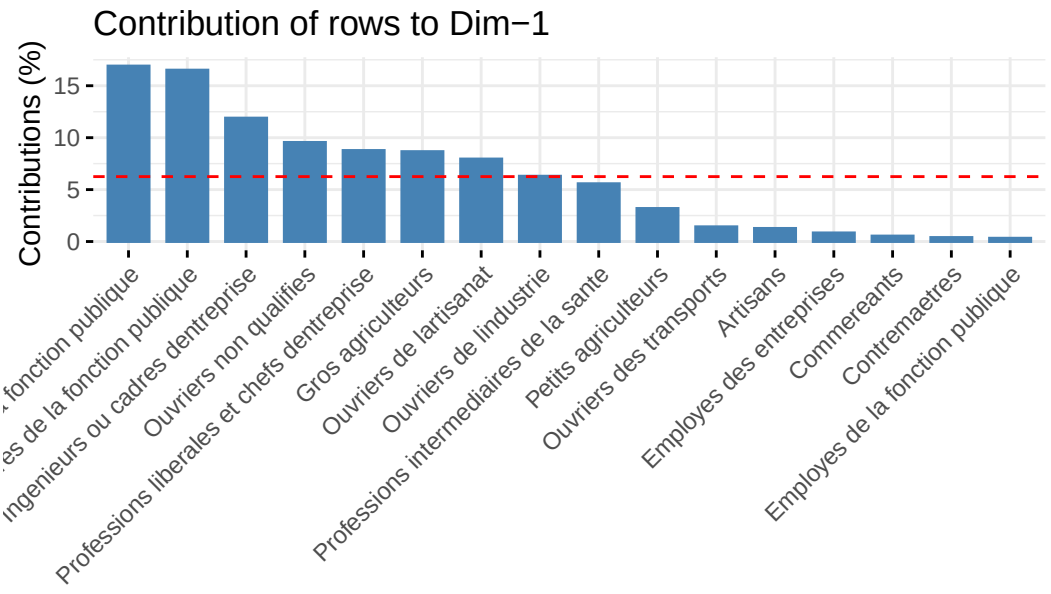
Ici, il y a 16 lignes, donc la contribution moyenne des profils-lignes est de 100/16=6.25. Il y a 14 colonnes, donc la contribution moyenne des profils-colonnes est de 100/14=7.1.

Interprétation de l'axe 1

La fonction fviz_contrib permet de visualiser facilement les contributions des modalités sur les axes.

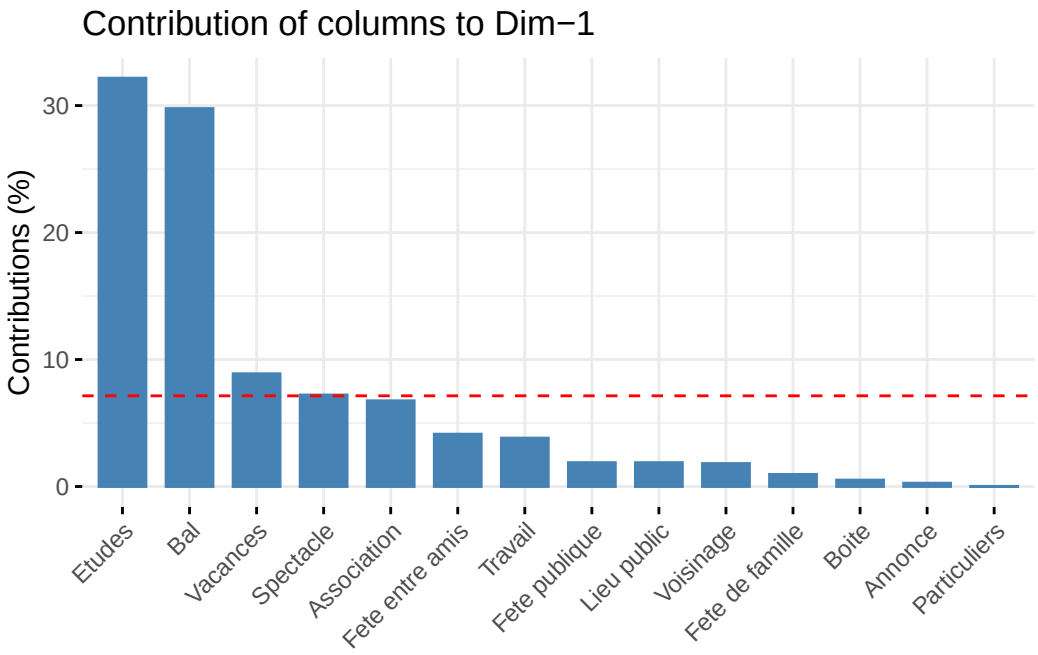
Sur l'axe 1, on interprète prioritairement les profils-lignes des cadres et professions intermédiaires de la fonction publique, des ingénieurs et cadres ou professions d'entreprises, des ouvriers non qualifiés, des professions libérales, des gros agriculteurs, des ouvriers de l'artisanat et des ouvriers de l'industrie.

```
##Visualiser la contribution des lignes par rapport à la contribution moyenne
fviz_contrib(res.afc,choice="row",axes =1)
```



Sur ce même axe, on interprète prioritairement les profils-colonnes des études, du bal, des vacances et des associations (à la limite).

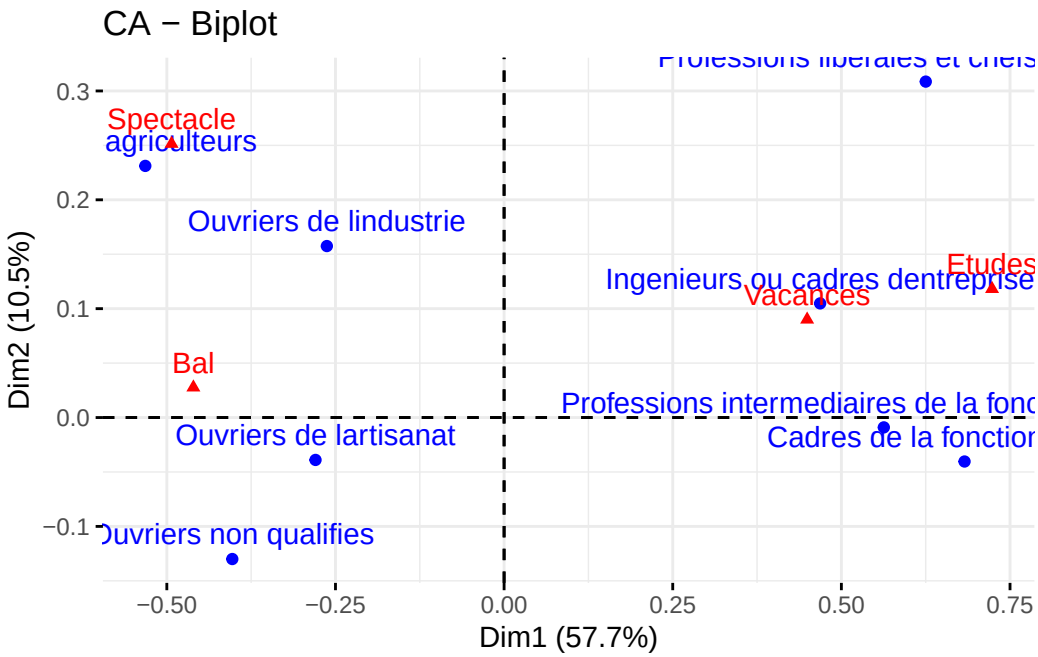
```
##Visualiser la contribution des lignes par rapport à la contribution moyenne
fviz_contrib(res.afc,choice="col",axes =1)
```



Voyons comment se structure alors le premier axe dans le premier plan factoriel (axes 1 et 2).

A gauche, on trouve les groupes sociaux du bas de la hiérarchie des PCS (les ouvriers, notons qu'il s'agit ici d'une ancienne nomenclature des PCS ! l'enquête date des années 1980) et les gros agriculteurs, qui ont des profils de lieu de rencontre similaires : au bal, au spectable. A droite, ce sont des classes supérieures et moyennes de la fonction publique, qui ont rencontré leur partenaire au cours de leurs études (ce sont aussi les groupes sociaux qui en moyenne ont fait les études les plus longues…) et en vacances (justement, ce sont les groupes sociaux qui partent le plus fréquemment en vacances…).

```
##Visualiser la contribution des lignes et colonnes qui ont une contribution supérieure à
rowcontrib1<-rownames(res.afc$row$contrib)[res.afc$row$contrib[,1]>100/16]

colcontrib1<-rownames(res.afc$col$contrib)[res.afc$col$contrib[,1]>100/14]

fviz_ca_biplot(res.afc,axes=c(1,2),select.row = list(name = rowcontrib1),select.col=list(n
```

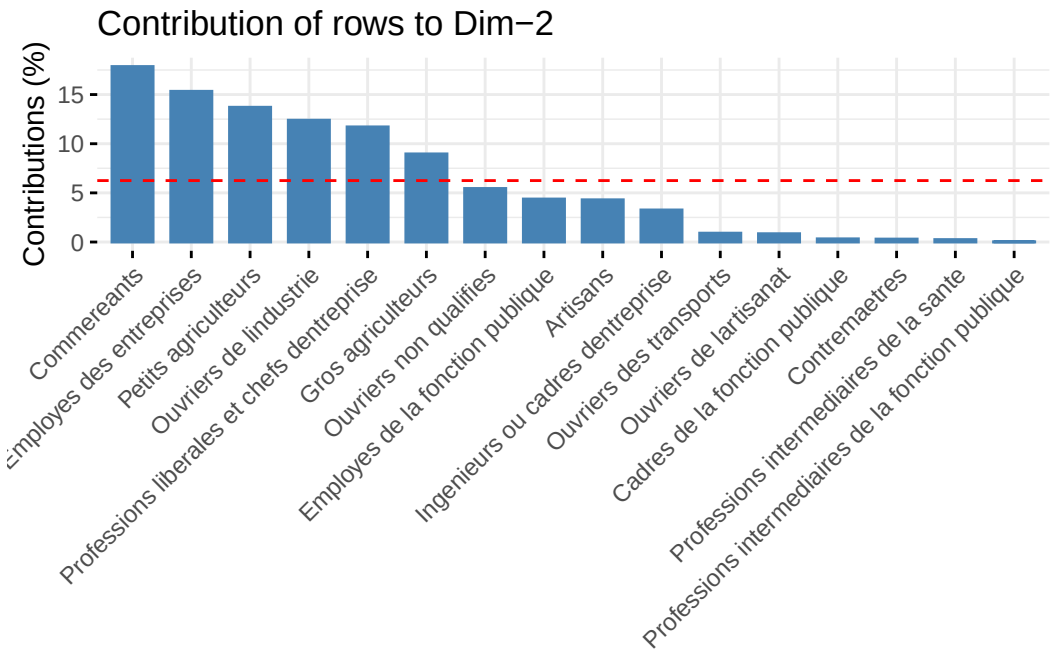


Interprétation des axes 2 et 3

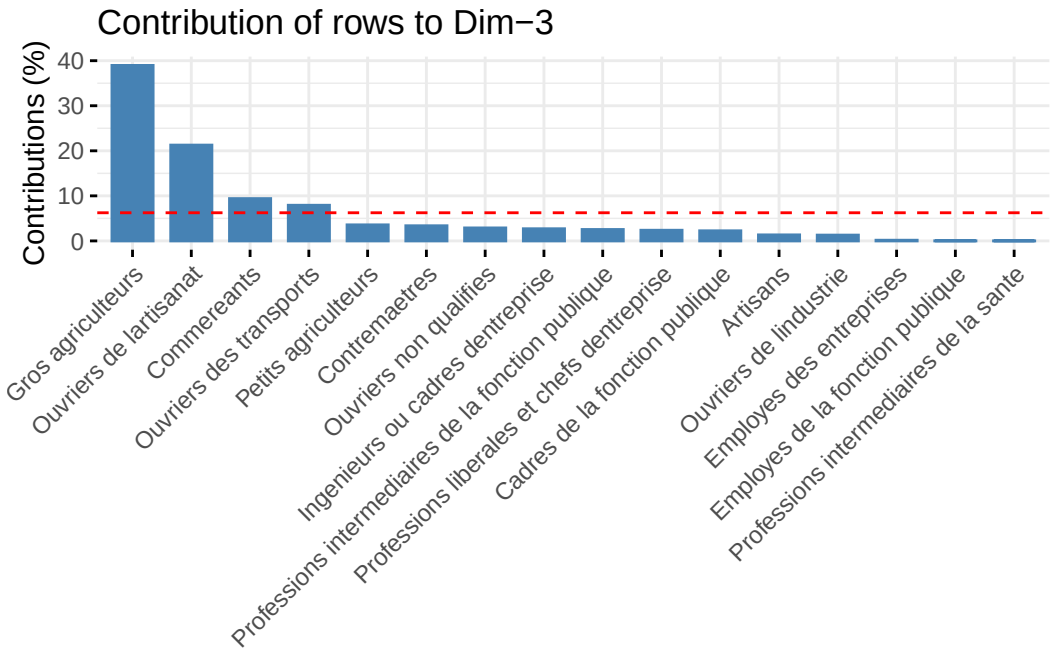
La même procédure peut être utilisée pour les axes 2 et 3.

Pour les profils-lignes :

```
##Visualiser la contribution des lignes par rapport à la contribution moyenne
fviz_contrib(res.afc,choice="row",axes =2)
```

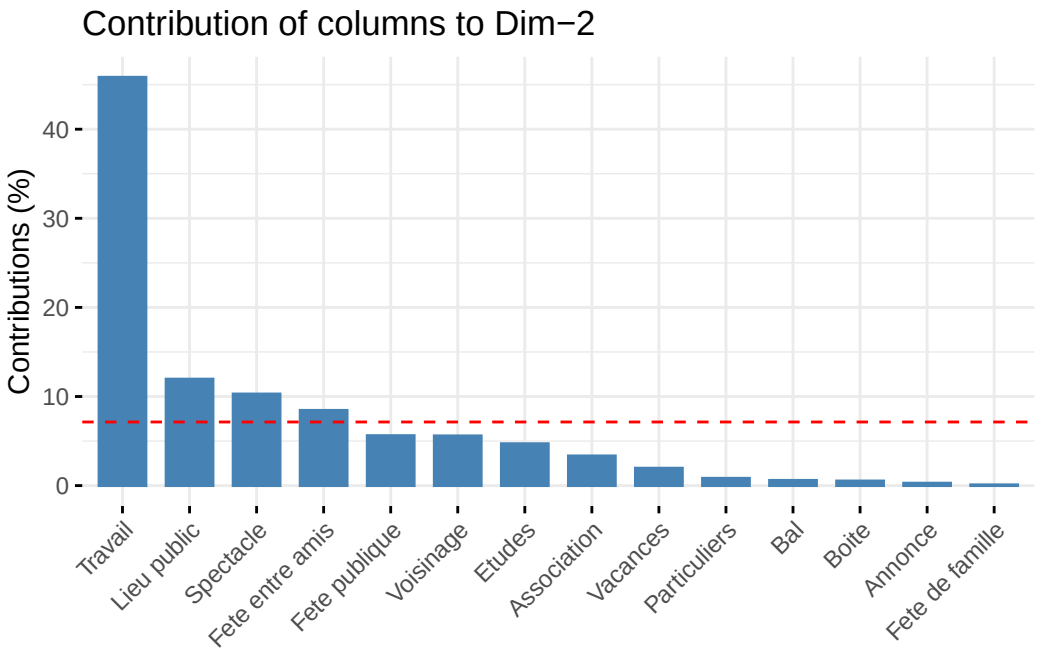



```
fviz_contrib(res.afc,choice="row",axes =3)
```

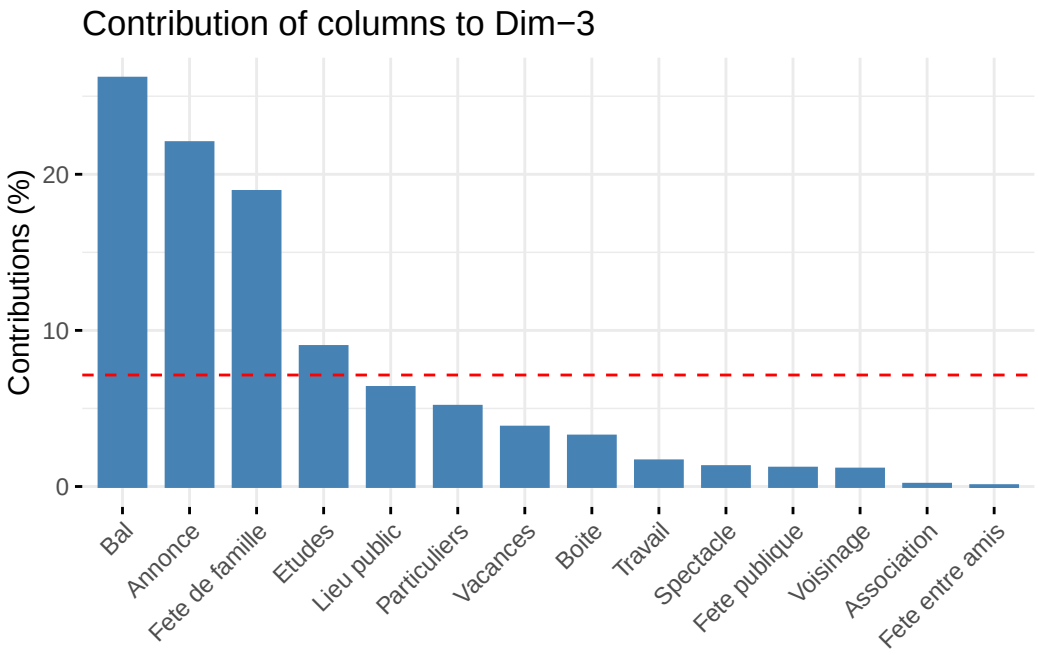


Pour les profils-colonnes :

```
##Visualiser la contribution des lignes par rapport à la contribution moyenne
fviz_contrib(res.afc,choice="col",axes =2)
```




```
fviz_contrib(res.afc,choice="col",axes =3)
```



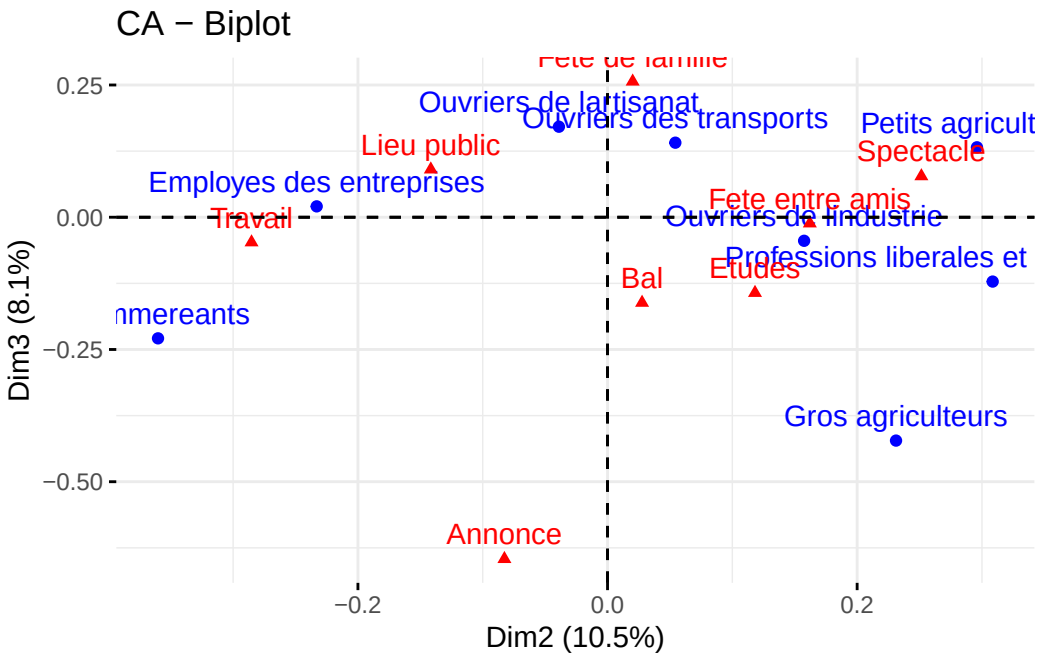
Voyons comment se structure alors le premier axe dans le plan factoriel des axes 2 et 3 qui présente, on l'a déjà dit des principes d'associations secondaires par rapport à l'association principale dénotée sur le premier axe de l'analyse.

L'axe 2 oppose les commerçants et les employés des entreprises qui rencontrent leur partenaire sur leur lieu de travail aux autres groupes sociaux. Il montre une spécificité du lieu de recontre pour ces deux groupes sociaux (par rapport aux autres professions où on se rencontre sur le temps de loisir).

L'axe 3 oppose les gros agriculteurs et dans une moindre mesure les commerçants qui rencontrent leur partenaire par le biais d'annonces par rapport aux autres groupes sociaux (en particulier les ouvriers). C'est un axe d'opposition entre la rencontre par un médium versus par des contacts sociaux.

```
##Visualiser la contribution des lignes et colonnes qui ont une contribution supérieure à
rowcontrib2<-rownames(res.afc$row$contrib)[res.afc$row$contrib[,2]>100/16]

colcontrib2<-rownames(res.afc$col$contrib)[res.afc$col$contrib[,2]>100/14]

rowcontrib3<-rownames(res.afc$row$contrib)[res.afc$row$contrib[,3]>100/16]

colcontrib3<-rownames(res.afc$col$contrib)[res.afc$col$contrib[,3]>100/14]
```

```
rowcontrib23<-c(rowcontrib2,rowcontrib3)
colcontrib23<-c(colcontrib2,colcontrib3)

fviz_ca_biplot(res.afc,axes=c(2,3),select.row = list(name = rowcontrib23),select.col=list(
```



Bien sur, rien n'empêche d'utiliser également la fonction explor() pour étudier les résultats de l'analyse !

On peut vouloir également mettre en relation nos interprétations en calculant les probabilités conjointes, les pourcentages en lignes et en colonne sur notre tableau de contingence. On écrira :

```
mytable<-as.table(as.matrix(TD4))
prop.table(mytable)*100
# cell percentages
prop.table(mytable, 1)*100
# column percentages
prop.table(mytable, 2)*100
# column percentages
```